START ~10

ACLESS DATABASE OF VECTORS OF AUTOMATIC TRANSCRIPTIONS OF DOCUMENTS ~20

TRUNCATE VECTORS TO REMOVE TERMS ~30

WEIGHT TRUNCATED VECTORS ~40

ADD TERMS NOT RECOGNIZED TO WEIGHTED, TRUNCATED VECTORS ~50

ARE OTHER VECTORS TO BE TRUNCATED ? ~60

YES

NO

70

COMPARE RETRIEVAL EFFECTIVENESS OF TRUNCATED, WEIGHTED VECTORS TO ORIGINAL VECTORS

STOP

Fig. 1

| Code | Provided By | WER |
|---|---|---|
| Human | NIST | 0% |
| CUHTK-S1 | Cambridge University | 24.8% |
| Dragon98-S1 | Dragon Systems | 29.8% |
| ATT-S1 | AT&T Labs | 31.0% |
| NIST-B1 | Carnegie Melon (CMU) | 34.1% |
| SHEF-S1 | Sheffield University | 36.8% |
| NIST-B2 | Carnegie Mellon (CMU) | 46.9% |
| DERASRU-S2 | DERA | 61.5% |
| DERASRU-S1 | DERA | 66.2% |

Table 1: Transcriptions used in this study.

Fig. 2
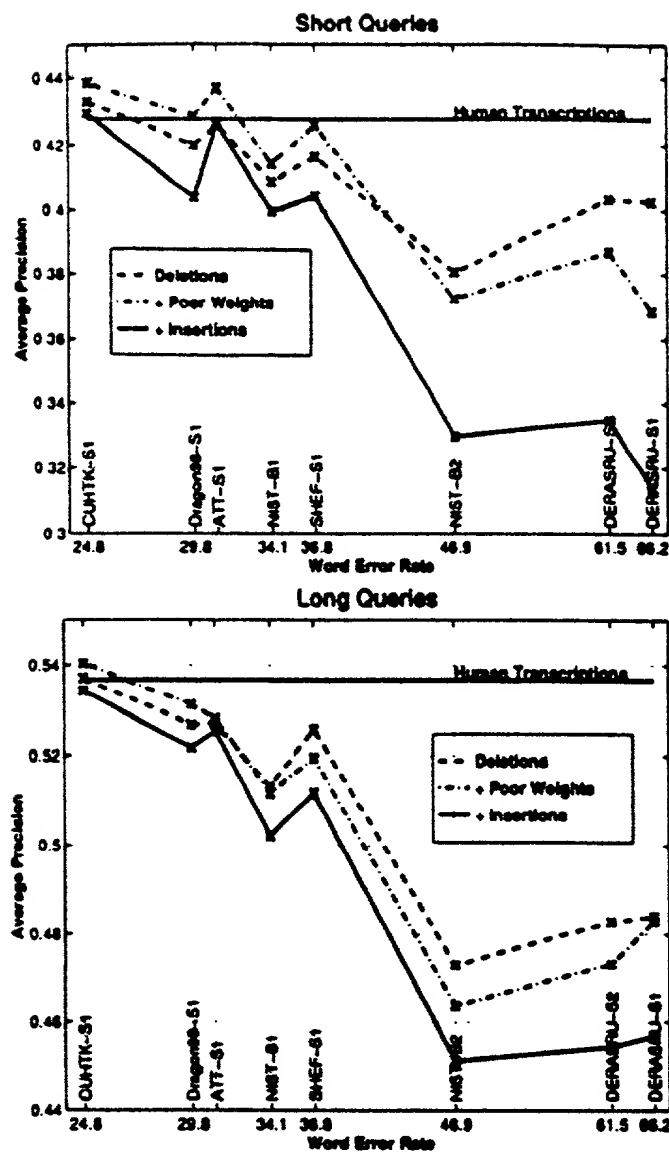
Figure 1: Losses due to different recognizer mistakes.

*Fig. 3*

| Degree of Expansion | $\alpha$ in Rocchio | | | |
|---|---|---|---|---|
| | 0.5 | 1.0 | 1.5 | 2.0 |
| 0% | 0.4739 | 0.4656 | 0.4591 | 0.4556 |
| 10% | 0.4898 | 0.4834 | 0.4756 | 0.4708 |
| 20% | 0.4938 | 0.4866 | 0.4813 | 0.4763 |
| 50% | 0.4990 | 0.4977 | 0.4898 | 0.4800 |
| 100% | 0.4992 | **0.4995** | 0.4952 | 0.4892 |
| 200% | 0.4943 | 0.4974 | 0.4923 | 0.4858 |

Table 2: Parameter exploration for NIST-B1 (short queries).

Fig. 4
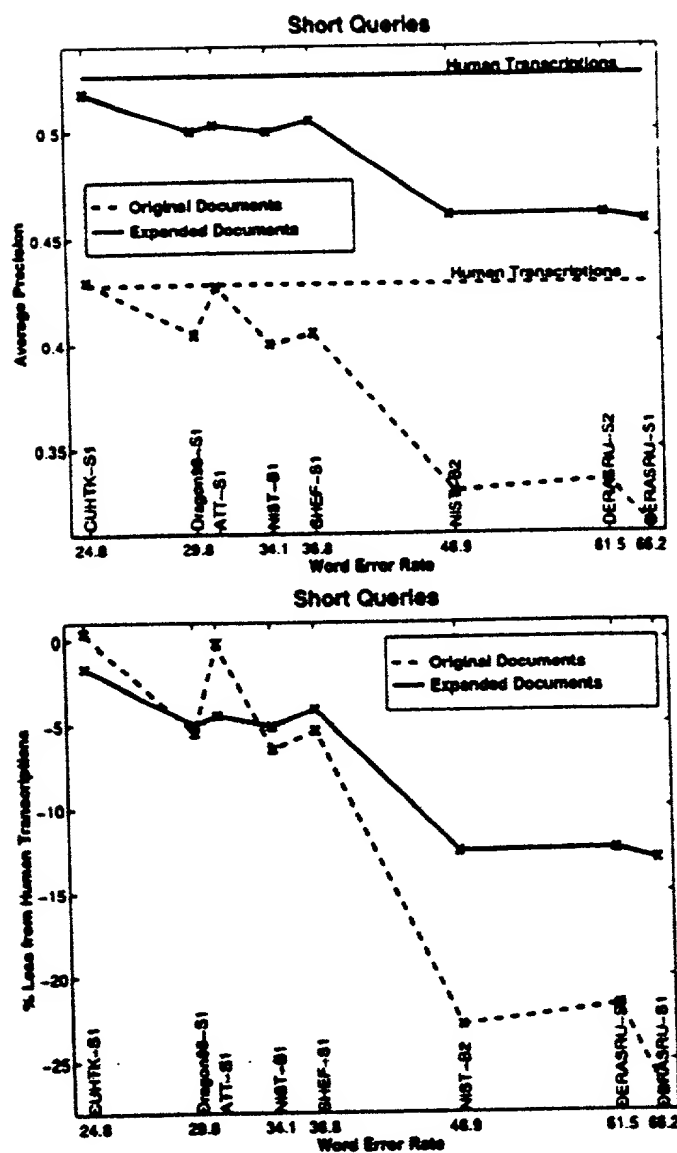
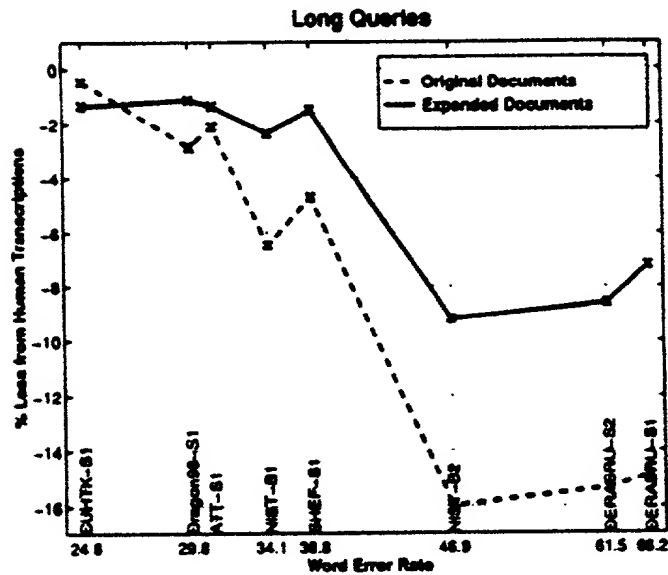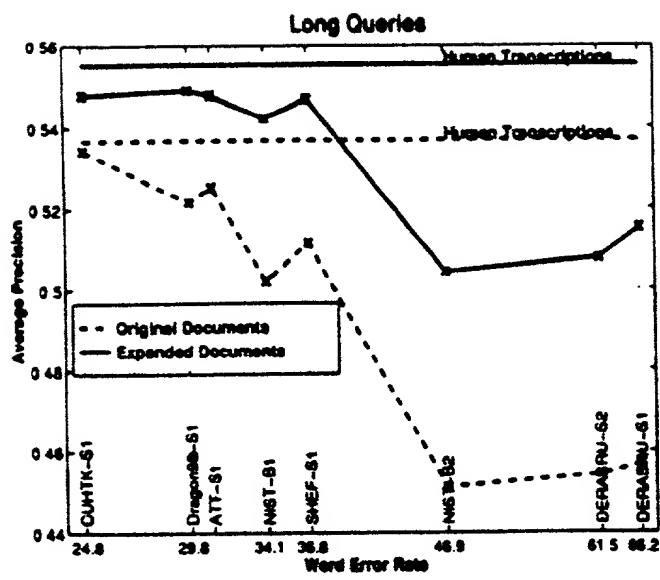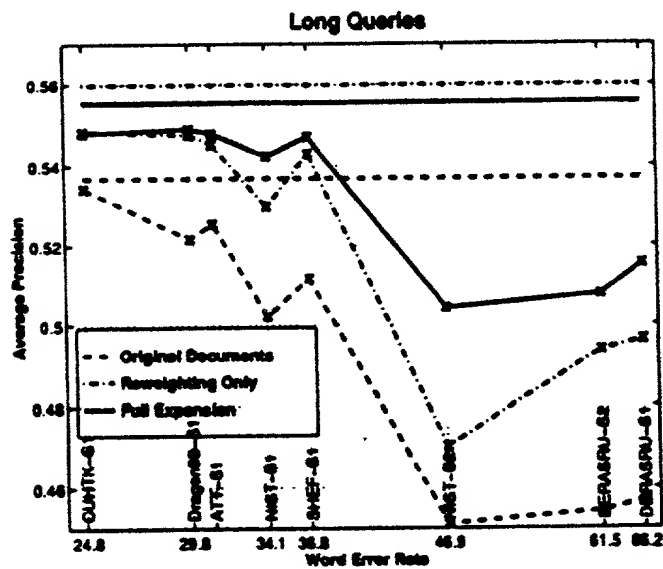Figure 2: Document expansion for short queries.

*Fig. 5*

Fig. 6

**Short Queries**

- - - Original Documents
- · - Reweighting Only
—— Full Expansion

Average Precision

Word Error Rate

**Long Queries**

- - - Original Documents
- · - Reweighting Only
—— Full Expansion

Average Precision

Word Error Rate

*Fig. 7*